Les 2. Analyse van één variabele

Onderzoekstechnieken

Jens Buysse **Wim De Bruyn** **Bert Van Vreckem**

AJ 2018-2019

HO GENT

What’s on the menu today?

Beschrijvende statistiek
- Centrummaten
- Spreidingsmaten

Grafieken
- Eenvoudige grafieken
- Interpretatie van grafieken

HO GENT

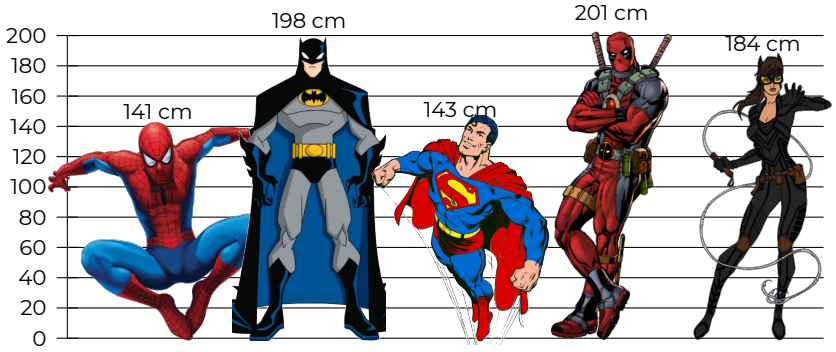
Stel, je wil een vriendengroep analyseren

Vragen die je kan stellen:

- Wat is de grootte van mijn vrienden?
- Hoeveel wegen mijn vrienden?
- Hoe veilig maken ze hun woonomgeving?
- Hebben ze familie?
- ...

HO GENT

Hoe groot zijn mijn vrienden?



HO GENT

Beschrijvende statistiek

HO
GENT

Centrummaten

Welke waarde is representatief voor de hele groep?

HO
GENT

Gemiddelde (of *mean*, *average*)

Het gemiddelde is de som van alle waarden gedeeld door het aantal waarden

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184

HO GENT

Gemiddelde (of *mean*, *average*)

Vraag 1 Wat gebeurt er als Kabouter Wesley (10cm) er bij komt?

Vraag 2 Het gemiddelde van 15 cijfers is 12. Welk nummer moeten we aan de rij van cijfers toevoegen om een gemiddelde van 13 te komen?

HO GENT

Mediaan (*median*)

Om de mediaan te vinden, sorteer de waarden en kies het middelste getal

- Oneven aantal getallen: geen probleem
- Even aantal getallen: gemiddelde van de middelste twee

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184

HO GENT

Mediaan (*median*)

Vraag 1 Wat gebeurt er nu als Kabouter Wesley er bij komt?

Vraag 2 Gegeven het totaal aantal mensen gered door Batman de laatste 8 jaar, wat is de mediaan?

4	7	11	16	20	22	25	26

HO GENT

Modus (*mode*)

De modus is het vaakst voorkomende getal in een reeks getallen

Totaal aantal mensen gered door Superman de laatste 15 jaar:

3	7	5	13	20	23	39	23	40	23	14	12	56	23	29

Totaal aantal mensen gered door Batman de laatste 8 jaar:

4	7	11	16	20	22	25	26

HO GENT

Spreidingsmaten

Hoe groot zijn de onderlinge verschillen binnen de groep?

HO GENT

Bereik (*range*)

Het bereik van een reeks getallen is de absolute waarde van het verschil tussen het grootste en kleinste getal in de reeks.

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184

HO GENT

Kwartielen (*quartiles*)

De kwartielen van een gesorteerde reeks getallen zijn de waarden die de lijst in vier gelijke delen verdeelt. Elk deel vormt dus een kwart van de dataset. Men spreekt van een eerste, tweede en derde kwartiel, genoteerd als resp, Q_1, Q_2, Q_3

Totaal aantal mensen gered door Superman de laatste 15 jaar:

3	7	5	13	20	23	39	23	40	23	14	12	56	23	29

HO GENT

Variantie en standaardafwijking

Eng. *variance*, resp. *standard deviation*

De variantie is het gemiddelde gekwadrateerde verschil tussen de elementen van de dataset en zijn gemiddelde

De standaardafwijking is de vierkantswortel van de variantie

x_1	x_2	x_3	x_4	x_5
141	198	143	201	184

HO GENT

Eigenschappen standaardafwijking

- Kan de standaardafwijking negatief zijn?

HO GENT

Eigenschappen standaardafwijking

- Kan de standaardafwijking negatief zijn?
- Wat is de kleinst mogelijke waarde? Wat duidt dit aan?

HO GENT

Eigenschappen standaardafwijking

- Kan de standaardafwijking negatief zijn?
- Wat is de kleinst mogelijke waarde? Wat duidt dit aan?
- Wat is de invloed van uitschieters op de standaardafwijking?

HO GENT

Eigenschappen standaardafwijking

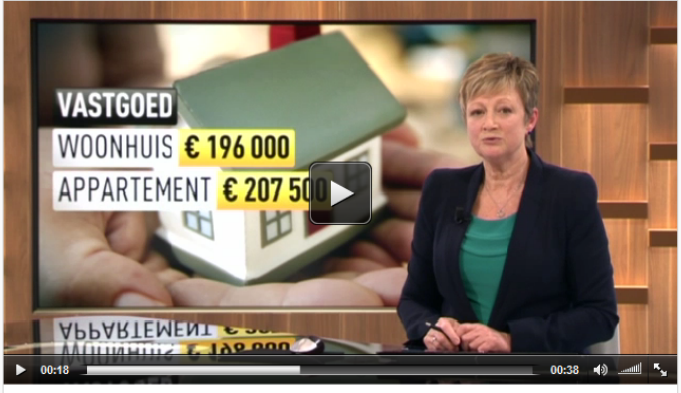
- Kan de standaardafwijking negatief zijn?
- Wat is de kleinst mogelijke waarde? Wat duidt dit aan?
- Wat is de invloed van uitschieters op de standaardafwijking?
- In welke eenheden staat de standaarddeviatie?

HO GENT

Eigenschappen standaardafwijking

- Kan de standaardafwijking negatief zijn?
- Wat is de kleinst mogelijke waarde? Wat duidt dit aan?
- Wat is de invloed van uitschieters op de standaardafwijking?
- In welke eenheden staat de standaarddeviatie?
- Hoe interpreteer je de standaardafwijking in combinatie met het gemiddelde?

HO GENT



Eigen dak boven hoofd wordt steeds duurder

Andere afleveringen

Het journaal 1 - 21/02/14

Het journaal L - 20/02/14

Het journaal 1 - 20/02/14

Het journaal L - 19/02/14

Onthou dit!

Enkel een centrummaat opgeven is nooit voldoende!

- Wat is de spreiding?
- Hoe is de data verdeeld? Normale verdeling?
- Is de groep voldoende homogeen?

HO GENT

Grafieken

HO
GENT

Cirkeldiagram (*pie chart*)

Waarom herkent niemand Superman?

HO GENT

Cirkeldiagram

Voordelen:

- Met percentages rond 20% kan je makkelijk verduidelijken t.o.v. volledige dataset

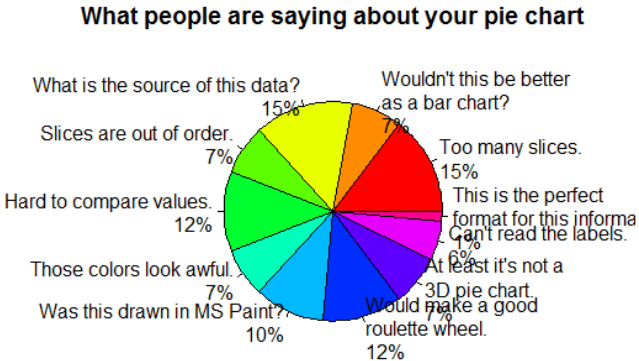
Nadelen:

- Vergelijken op basis van hoek, terwijl lengte intuïtiever is
- Figuur onduidelijk bij veel categorieën

HO GENT

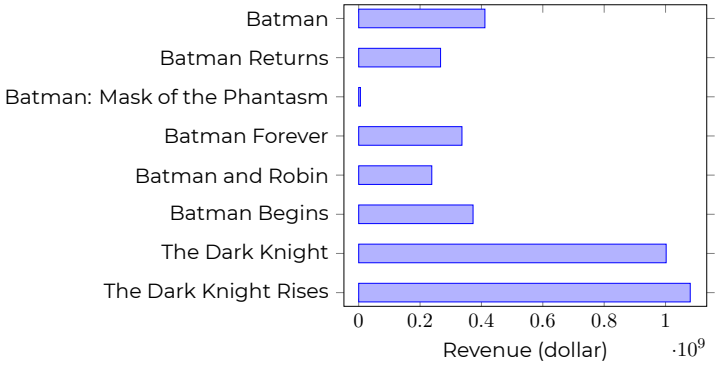
Cirkeldiagram

Vermijd het gebruik van cirkeldiagrammen!



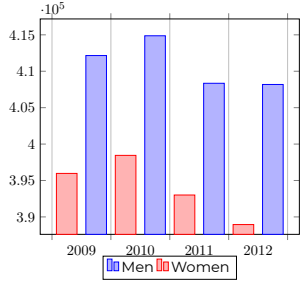
HO GENT

Staafdiagram (*bar chart*)



HO GENT

Staafdiagram

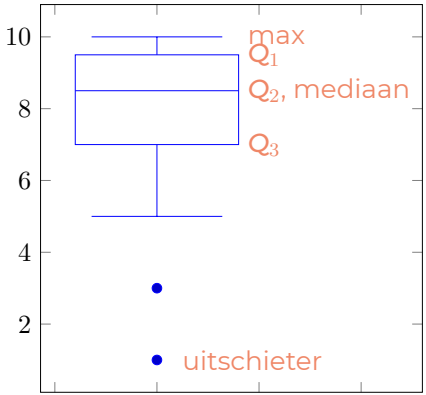


Voordelen:

- Makkelijker vergelijken van categorieën
- Per categorie meerdere “bars” mogelijk

HO GENT

Boxplot

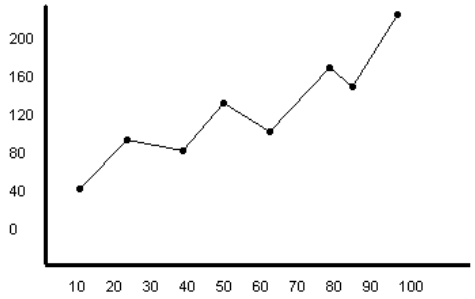


Voordeel: snelle manier om data te inspecteren en verschillende datasets te vergelijken

HO GENT

Data-ambiguïteit

= Niet aanduiden wat de data betekent.



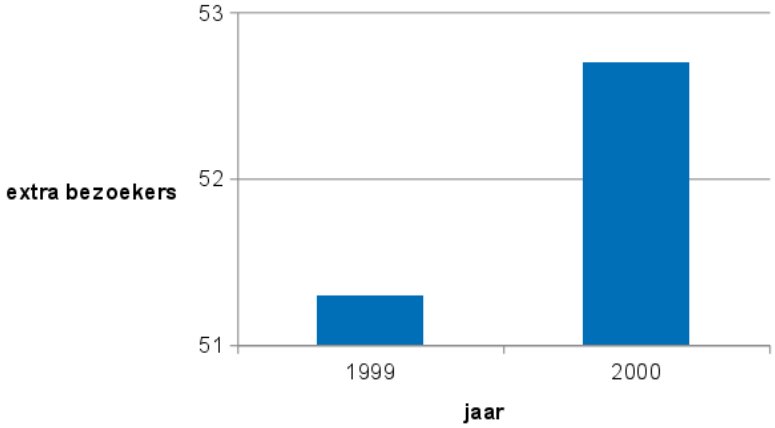
Tips:

- Benoem je assen
- Geef een duidelijke titel
- Benoem de meeteenheid (en evt. grootorde)
- Voeg een bijschrift toe met uitleg over de grafiek

HO GENT

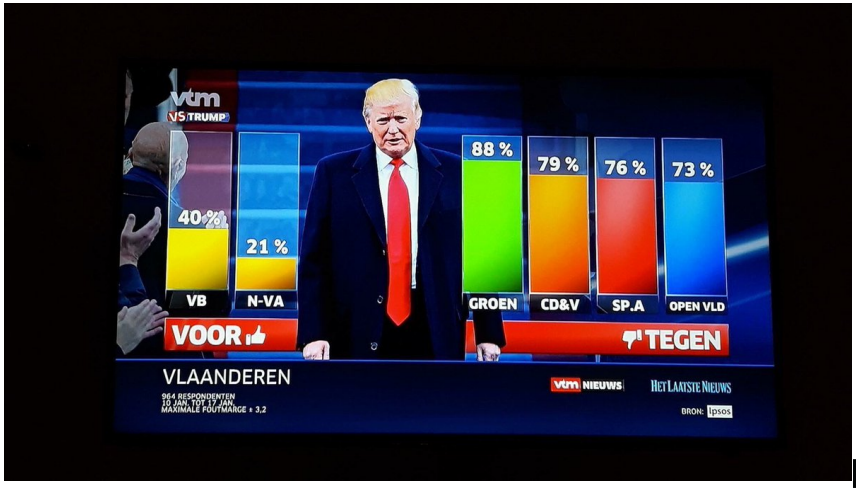
Data distortion

= Verkeerde conclusies laten trekken uit een grafische voorstelling



HO GENT

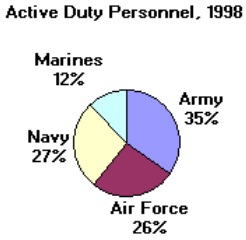
Data distortion

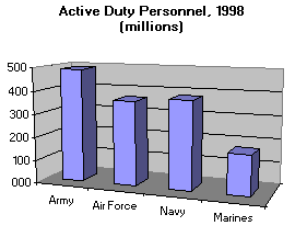


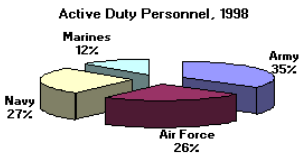
HO GENT

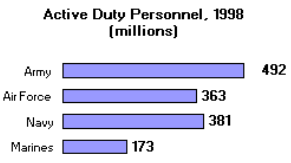
Data distraction

- Vermijd toeters en bellen in je grafieken
- Minimaliseer inkt tot data ratio



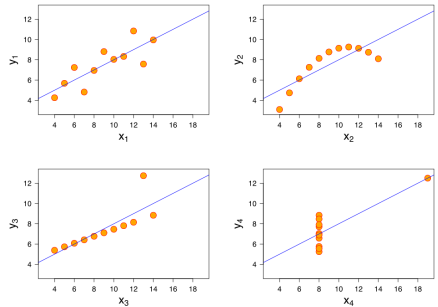






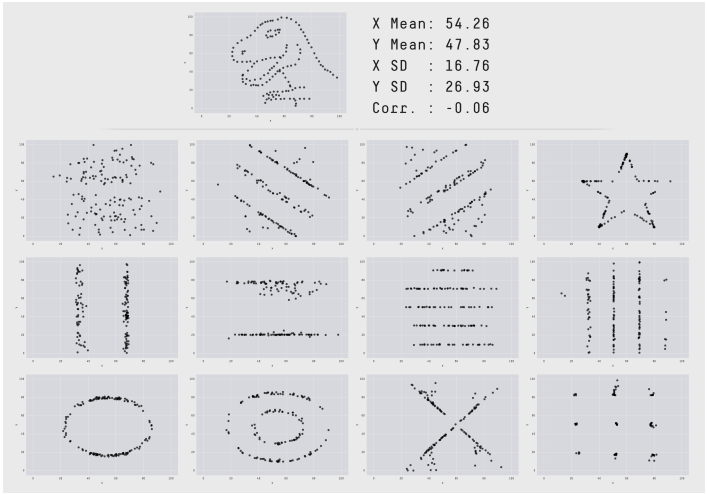
HO GENT

Anscombe’s Quartet



Vier verschillende datasets met dezelfde statistische eigenschappen.
Deze tonen het belang aan van data-visualisatie.

HO
GENT



“The datasaurus dozen” (Bron: https://www.autodeskresearch.com/publications/samestats)